DOCUMENT RESUME

ED 468 961 TM 034 449

AUTHOR Lee, Yong-Won

TITLE Score Reliability of a Test Composed of Passage-Based Testlets: A Generalizability Theory Perspective.

PUB DATE 2002-08-00

NOTE 24p.; Paper presented at the International Conference of the Korea English Education Society (1st, Chungbuk, Korea, August 2-3, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS

DESCRIPTORS English (Second Language); Foreign Countries; *Generalizability Theory; *High School Students; High Schools; Item Response Theory; Reading Tests; *Reliability; *Scores; Simulation

IDENTIFIERS *Item Dependence; South Korea

ABSTRACT

The purpose of this study was to investigate the impact of local item dependence (LID) in passage-based testlets on the test score reliability of an English as a Foreign Language (EFL) reading comprehension test from the perspective of generalizability (G) theory. Definitions and causes of LID in passage-based testlets are reviewed within the frameworks of both classical test theory (CTT) and item response theory (IRT) along with three reliability estimation procedures for testlet-based tests. Scored item responses of 1,072 Korean high school students to a 40-item, 10-passage EFL reading comprehension test were analyzed along with simulated data generated based on the real data by using the computer program GENOVA (G. Crick and R. Brennan, 1983). It was found that the item alpha coefficient from the p x I design was higher than the passage alpha coefficient from the p x T and the generalizability coefficient from the p x (I:H) designs in real data analyses. In simulated data analyses, however, the G-coefficients from the p x I, p x T, and P x (I:H) designed were very close because no LID was assumed within item blocks. Subsequent analyses of the fixed test length scenarios have also revealed that score reliability could be maximized by increasing the number of passages rather than the number of items nested within passages. (Contains 1 table, 3 figures, and 33 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 961

Score Reliability of a Test Composed of Passage-Based Testlets: A Generalizability Theory Perspective

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Y-W. Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee, Ph. D.

Center for Validity Research
Division of Research and Development
Educational Testing Service (ETS)
Princeton, NJ 08541, USA

Phone: (609)-734-1348
Fax: (609)-734-1755
Email: ylee@ets.org

Paper presented at the first international conference of Korea English Education Society (KEES) held at Korea National University of Education (KNUE) on August 2-3, 2002

TM034449

BEST COPY AVAILABLE

Abstract

The purpose of a study reported in this paper is to investigate the impact of *local item dependence* (LID) in passage-based testlets on the test score reliability of an EFL reading comprehension test from the perspective of *generalizability* (G) *theory*. Definitions and causes of LID in passage-based testlets are reviewed within the frameworks of both classical test theory (CTT) and item response theory (IRT) along with three reliability estimation procedures for testlet-based tests. Scored item responses of 1,072 Korean high school students to a 40-item, 10-passage EFL (English as a Foreign Language) reading comprehension test were analyzed along with simulated data generated based on the real data by using the computer program GENOVA (Crick & Brennan, 1983). It was found that the item α coefficient from the p x I design were higher than the passage α coefficient from the p x T and the generalizability ($E\rho^2$) coefficient from the p x (I:H) designs in real data analyses. In simulated data analyses, however, the G-coefficients from the p x I, p x T, and p x (I:H) designs were very close due to no LID assumed within item blocks. Subsequent analyses of the fixed test length scenarios have also revealed that score reliability could be maximized by increasing the number of passages rather than the number of items nested within passages.

Introduction

In reading comprehension assessment, a reading passage is often used as stimulus material for a series of related comprehension items (or tasks). These items can be positioned after the passage, as in a traditional multiple-choice or a short-answer constructed-response item test, or embedded in the passage itself, as in *cloze* or *c*-tests (Alderson, 2000; Klein-Braley, 1985). When multiple items are created based a common passage, it constitutes a viable assessment format (Haladyna, 1992), in which assessment context is provided for related comprehension tasks by the text. Such a format also makes it possible to measure examinees' comprehension of the material from various perspectives and at the same time is cost-effective both for item writers and examinees (Mehrens & Lehman, 1978; Thissen, Steinberg, & Mooney, 1989). To put it another way, item writers may not have to spend too much time preparing a lot of passages to generate a single question per passage, and the examinees also would not need to read long passages to answer only one question from each passage.

Such clusters of related items based on the same stimulus material are often called *testlets* in the measurement literature, for they can be used units of scoring and analysis as well as building blocks of the test (Wainer & Kiely, 1987; Wainer & Lewis, 1990). However, such passage-based testlets could pose a problem for item-based test analysis and construction within the framework of contemporary psychometric models. It is assumed implicitly in classical test theory (CTT) and explicitly in item response theory (IRT) that test items should be *locally independent* (Henning, 1989; Sireci, Thissen, & Wainer, 1991; Yen, 1984, 1993). When a series of items share a common reading passage, it is very likely that the *local item independence* (LII) assumption is violated among those items. In consequence, inaccurate estimation of examinee abilities, and item information can be obtained, especially when IRT models assuming LII are applied to the tests that include sets of *locally-dependent* items (Ferrara, Huynh & Michaels, 1999; Wainer & Thissen, 1996; Yen, 1993). It has been also known that when the item-based reliability estimation method based on CTT is applied to such tests, test reliabilities could be overestimated as a consequence of ignoring this LID structure in the test data.

Within the framework of IRT, one intriguing approach to this problem is to group together items that are *locally-dependent* and treat them either (1) as a single polytomous item or a *testlet* (Andrich, 1985; Thissen et al., 1989; Wilson & Adams, 1995) or (2) as a cluster of items sharing an additional item parameter (Wainer & Wang, 2000). Analogous solutions have been proposed in the frameworks of CTT and generalizability (G) theory to appropriately model the impact of LID in reliability estimation: (1) a passage alpha (α) based on polytomous testlet scores (Sireci et al., 1991; Wainer & Thissen, 1996) and (2) a generalizability coefficient based on separate modeling of passage and item effects through G-theory analysis (Lee, Brennan & Frisbie, 2001; Lee & Frisbie, 1999). The purpose of a study reported in this paper is to investigate the impact of passage-related *local item dependence* (LID) and the number of passages and items per passage on test score reliability of an EFL reading comprehension test from the perspective of *generalizability* (G) *theory*. In this paper, the author first describes how these passage-

based testlets can be handled in the three different reliability estimation methods (*i.e.*, item α, passage α, and G-theory-based *generalizability coefficient*) and present the results of real and simulated data analyses using these three different approaches.

Reliability Estimation for Passage-Based Tests

Local Item Dependence (LID) in Passage-Based Testlets

Local item dependence (LID) can be defined in both IRT and CTT (Henning, 1989; Lord & Novick, 1968; Yen, 1993). In IRT, the assumption of local item independence (LII) stipulates that responses to two different items should be statistically independent of each other for examinees at a fixed ability/trait level (Hambleton & Swaminathan, 1985). In CTT, it is assumed that errors of measurement are statistically independent among different items in the computation of standardized item alpha (Cureton, 1965; Suen, 1990) and thus not correlated over items in a test, given the examinee's true score (Yen, 1993).

In reading comprehension tests, most of the items sharing a passage are basically aimed at assessing examinees' comprehension of the related informational elements in the same text. Examinees' performance on one item may be thus more highly correlated to their performance on another item in the same passage than on any item in a different passage. In an IRT sense, however, not all of the excessive interitem covariation within the passage may contribute to a "a common latent trait" ((θ) underlying a whole collection of items across different texts in the test. The extra item dependence within the passage, which is unaccounted for by this common latent trait, is very likely to lead to "correlated residuals" or LID. It may also indicate that there is a secondary test dimension (s) tapped by these items, but unaccounted for by the IRT *theta* (θ).

In the computation of standard item alpha in the CTT framework, the average of the total inter-item correlations across all items in the test is used to compute the item alpha for the full test length by applying the Spearman-Brown Prophecy Formula. However, the average of *within-passage* inter-item correlations are likely to be systematically higher than that of *between-passage* inter-item correlations. It may be indicative of some extra interitem covariation within passages unattributable to examinees' "true scores" generalizable across passages in the test. Nevertheless, the average of the total inter-item correlations would be a weighted average of these *within-passage* and *between-passage* averages, and such extra item dependence within passages may result in an inflated estimate of item alpha and lead to the violation of the assumption of "uncorrelated error among items".

In principle, if the content of one item provides clues to the answer of another item in a test, LID occurs (Hambleton & Swaminathan, 1985). Yen (1993) has provided a list of potential causes of LID including the sharing of a common passage, content, knowledge, item chaining, speededness, fatigue, practice effects, item or response format, and so forth. In EFL reading comprehension tests, the physical layout of the test booklet

could also become a potential source of LID for some items (Muraki & Lee, 2001) as well as overlapped information between the options of one item and the stem of another item within a passage (Lee, *in press*). Chen and Thissen (1997) have made a distinction between "the underlying LID" stemming from the contents of items and "the surface LID" often found in a series of not-reached items at the end of the test.

Score Reliability Estimation for Passage-Based Testlets

Item Alpha (α) Based on Dichotomously-Scored Individual Items. A Cronbach (1951) alpha (or KR-20) based on this approach is equivalent to a *generalizability* coefficient ($E\rho^2$) based on a single-facet design with items as a random facet (p x I) in the G-theory framework (Brennan, 1992, Suen, 1990). The p x I design ignores the structure of passage-based testlets and treat the passages as a "hidden fixed facet" (Lee & Frisbie, 1999). Previous studies have shown that the reliability of test scores composed of passage-based testlets could be overestimated, when the conventional item-based reliability estimation methods (*i.e.*, standardized item α) are used (Sireci, Thissen & Wainer, 1991; Wainer, 1996; Wainer & Thissen, 1996; Yen, 1993). These studies have shown that the inflated portion of score reliability can be removed by using the testlet scores as the unit of analysis instead of items.

Passage Alpha (α) Based on Polytomously-Scored Passage-Based Testlets. A Cronbach alpha based on this approach is equivalent to a *generalizability* ($E\rho^2$) coefficient based on a single-facet design with passage-based testlets as a random facet (p x T). This approach involves treating a set of related (or locally-dependent) items as a single polytomous item or *a testlet* (Wainer & Lewis, 1990). This approach uses polytomously-scored *testlets* (or passages) as the main unit of analysis to remove the impact of *local item dependence* on test score reliability. It has been known that when there is LID within testlets, the item α is higher than the passage α and the difference between the two is the inflated portion of the α-coefficient due to LID within the passages (Wainer & Lukhele, 1997; Wainer, Sireci, & Thissen, 1991; Wainer & Thissen, 1996). However, it has been argued that using testlet scores could lead to some loss of information and lower reliability estimates as well relative to a more precise method (Lee & Frisbie, 1999, Yen, 1993). Recently, it has been also demonstrated that the *generalizability* approach to reliability estimation could solve such a problem (Lee & Frisbie, 1999).

Generalizability ($E\rho^2$) Coefficient Based on Dichotomously-Scored Items Nestsed within Passages. A univariate single facet design with persons crossed with items nested within passages (p x (i:h)) might be the most appropriate design for a test composed passage-based testlets (Lee, Brennan & Frisbie, 2001). This approach treats persons (*p*) as objects of measurement and items (*i*) and passages (*h*) as random facets and makes it possible to differentiate the various sources of measurement error variance. In the p x (i:h) design for the univariate analysis, all of the 5 main and interaction variance components (σ_p^2, σ_h^2, $\sigma_{i:h}^2$, σ_{ph}^2, $\sigma_{pi:h}^2$) can be estimated in the *generalizability* (G) *study*. In *the decision* (D) *study*, two different kinds of score reliability coefficients can be

computed for different measurement scenarios, that is, a *generalizability* ($E\rho^2$) *coefficient* and a *dependability* (Φ) *index*. Assuming a balanced design in which the number of items is the same across passages, the $E\rho^2$ and Φ coefficients can be computed based on the p x (i:h) design as follows:

First, the relative error term (δ) and the *generalizability* coefficient ($E\rho^2$) can be defined, as in equation 1, which can be interpreted as the error variance and the reliability coefficient for norm-referenced score interpretation, respectively (Brennan, 1992; Suen, 1990).

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{ph}^2}{n_h} + \frac{\sigma_{pi:h}^2}{n_{ih}}} \quad \text{----------(1)}$$

Because the passage effect is treated as a hidden fixed facet in the p x I design, σ^2 *(pH)* is included in the universe score variance (Lee & Frisbie, 1999) rather than in the relative error variance (δ) for the item α.

Second, the absolute error (Δ) and the dependability index (Φ) can be computed, as in equation 2, which can be interpreted as the error variance and the score reliability coefficient for criterion-referenced score interpretation, respectively. When the mean score is applied as a cut-off score for admission and licensure purposes, the Φ coefficient is more appropriate.

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_h^2}{n_h} + \frac{\sigma_{i:h}^2}{n_{ih}} + \frac{\sigma_{ph}^2}{n_h} + \frac{\sigma_{pi:h}^2}{n_{ih}}} \quad \text{.....(2)}$$

Research Questions

1. How serious would the reliability overestimation be, when the item-based reliability estimation method is applied to tests made up of *passage-based testlets*?

2. How robust is the testlet-based reliability estimation method, compared to the G-theory-based method?

3. What is the impact of the numbers of testlets and items per testlet on the score reliability of tests made up of *passage-based testlets*?

Method

Data

Both real and simulated data were used. Real data for this study were scored item

responses of 1,072 Korean EFL high school examinees to a 40-item, 10-passage EFL reading comprehension test. There were 4 items given in each of the 10 passages. The test was administered at five high schools in a metropolitan area including Seoul during the fall of 1997. There were 388 tenth graders and 684 eleventh graders. Eleventh graders have taken one more year of English than the tenth graders. The number of participants at each school ranged from 99 to 278. Of 1072 examinees, 679 were male examinees and 393, female. In previous research (Lee, in press), passage-related LID was confirmed across all 10 passages used in this study through Q_3 analyses.

Simulated data were used to provide a baseline for a no LID situation. Item parameters obtained from the MULTILOG (Thissen, 1991) calibration of the real data were used to generate binary item responses for a hypothetical 40-item test, with local item independence assumed for all 40 items. The three-parameter logistic (3PL) IRT model was used to generate simulated data by using the computer program RESGEN (Muraki, 2000). The following specifications were used to generate the simulated data set: 10 item blocks, 4 items in each block, and a normal distribution of thetas (θ), with a mean of 0.0 and a standard deviation of 1.0. (See Lee (in press) for more details about the real and simulated data).

Analysis

The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components and $E\rho^2$ and Φ coefficients for the real and simulated data in the three G-study designs: p x i, p x t, and p x (i:h). D-studies were designed for the same universe of generalizations as the universe of admissible observations used in the G-study for all three designs (*i.e.*, p x I, p x T, and p x (I:H))[1].

First, a single-facet crossed design (p x i) with items as a random facet was employed in the G-study to estimate the variance components for the main and interaction effects separately for the real and simulated data sets. The "p x I" design was used in the D-study to estimate $E\rho^2$ coefficients based on dichotomously-scored individual items along with its Φ coefficient counterpart. As previously mentioned, an "item α" equals an $E\rho^2$ coefficient in a single-facet p x I design (Brennan, 1992; Suen, 1990). Multiple D-studies for computing $E\rho^2$ and Φ coefficients were carried out by varying the number of items from 1 to 40.

Second, a single-facet crossed design (p x t) with passage-based testlets as a random facet to estimate the variance components for the 10-testlets tests in a G-study. The "p x T" design was used in the D-study to estimate the $E\rho^2$ (passage α) and Φ coefficients based on polytomously-scored passage based testlets. A "passage α" also equals an $E\rho^2$ coefficient in this single facet p x T design. Multiple D-studies were carried

[1] In *generalizability theory* notation, measurement facets in a *generalizability* study (G-study) are identified by lowercase letters (e.g., "i" and "h"), but the facets in a *decision* study (D-study) are identified by uppercase letters (e.g., "I" and "H"). However, the person facet (p) is represented by a lower case letter in both G- and D- studies.

out by increasing the number of testlets from 1 to 10.

Third, a partially-nested two-facet design (p x (i:h) with items nested within passages was used in the G-study to estimate the variance components for the 40-item tests. A balanced nesting design was used, which assumes that there would be the same number of items (*e.g.*, 4 items per passage) nested within each of the 10 passages. Multiple D-studies (p x (I:H)) were conducted by: (1) increasing the total test length from 1 to 40 items (*i.e.*, increasing the number of passages, with the number of items nested within passages fixed to 4, as in the actual 40-item test); (2) increasing the number of passages, with the number of items nested within passages fixed to 1, 2, 3, and 4, respectively; and finally (3) manipulating the scenarios of items nested within passages for the total test length of 40 items.

For the first investigation condition of "increasing the total test length from 1 to 40 items", the $E\rho^2$ coefficients estimated from the p x I, p x T, and p x (I:H) designs for different test lengths were plotted together to make comparison possible among the three methods. For the second condition of "increasing the number of passages, with the number of items per passage fixed to 1, 2, and 3 as well as to 4 ", the $E\rho^2$ coefficients estimated only from the p x (I:H) designs based on the real and simulated data were plotted together. Finally, for the third condition of "the fixed total test lengths of 40 items", there were 8 possible item nesting scenarios for the p x (I:H) design, which included the scenarios of "40 items per passage for a single passage" ($40i{:}1p$), "20 items per passage for each of 2 passages" ($20i{:}2p$), "10 items per passage for each of 4 passages" ($10i{:}4p$), "8 items per passages for each of 5 passages" ($8i{:}5p$), "5 items per passage for each of 8 passages" ($5i{:}8p$), "4 items per passage for each 10 passages" ($4i{:}10p$), "2 items per passage for each of 20 passages" ($2i{:}20p$), and "1 item per passage for each of 40 passages" ($1i{:}40p$).

Results

Reliability estimates from multiple D-studies are presented in the following order: (1) increasing the total test length from 1 to 40 (p x I, p x T, and p x (I:H)); (2) increasing the number of passages from 1 to 10, with the number of items nested within passages fixed to 1, 2, 3, and 4, respectively (p x (I:H)); and (3) manipulating the scenarios of items nested within passages for the fixed total test length of 40 items (p x (I:H)).

Condition 1: Increasing the Total Test Length

Tables 1-a and 1-b shows the *generalizability coefficients* ($E\rho^2$) and the *dependability indices* (Φ) from the p x I, p x T, and p x (I:H) designs for different total test lengths. Because it was assumed that the number of items nested within each passage was fixed to 4 as in the actual 40-item test used in this study, 20 and 40 items in the p x I design would be equivalent to 5 and 10 passages-based *testlets* in the p x T and p x (I:H) designs in terms of test length, respectively. Figures 1-a and 1-b also graphically display

the $E\rho^2$ coefficients for various test lengths that were estimated from the real and simulated data sets.

Insert Table 1 and Figure 1 about here

The item α coefficients from the p x I design were higher than the passage α coefficients from the p x T design and the $E\rho^2$ coefficients from the p x (I:H)) design for different test lengths in the real data analysis. Nevertheless, the passage α and the $E\rho^2$ coefficients were very close and almost identical, as shown in Figure 1-a. For instance, the item-α coefficient for the full 40-item (or 10-testlet) test was about 0.89, whereas the passage α and the $E\rho^2$ coefficients were all 0.87. The difference between the first and the latter two methods was the smallest (0.02) for the full 40-item test scenario but the largest (0.06) for a 4-item (a single-passage) test. In other words, the difference became smaller, as the total number of items increased. The similar patterns were observed for the Φ coefficients for various test lengths.

In simulated (no LID) data analyses, the item α coefficients from the first design was also slightly higher than the passage α and the $E\rho^2$, but these three types of reliability coefficients were very close and had only negligible differences for different test lengths, as shown in Table 1-b and Figure 1-b. For instance, all three of the item α, the passage α, and the $E\rho^2$ coefficients for the full 40-item (or 10-testlet) test were approximately 0.89, when rounded off to the second decimal point. However, the difference between the first and the latter two designs was more apparent for shorter test lengths, which ranged from 0.00 to 0.02. The similar patterns were observed for the Φ coefficients for various test lengths.

Condition 2: Increasing the Number of Passages, with the "Number of Items Nested within Passages" Fixed to 1, 2, 3, and 4

Figure 2 shows how the $E\rho^2$ coefficients from the p x (I:H)) design would change, as the number of passages increased from 1 to 10 for four different scenarios of the fixed testlet lengths (*i.e.*, 1, 2, 3, and 4 items per passage). The four solid lines represent the $E\rho^2$ coefficients from the real data, whereas the four dotted lines represent those from the simulated data.

Insert Figure 2 about here

As shown in Figure 3, the dotted (simulated data) lines were consistently higher than the solid (real data) lines for the 2, 3, and 4 items per passage scenarios, although both dotted and solid lines were almost identical and laid on top of each other for the

"one-item-per- passage" scenario. It was also observed that the difference between the solid and dotted lines for the same scenario became larger, as the number of items per passage increased. It may indicate that there would be comparatively less gain in score reliability for increasing the number of items per passage in real (LID) than simulated (no LID) data.

Nevertheless, the slopes of the solid and dotted lines passage were the same in both the real and simulated data analyses for different numbers of passages, which suggests that the rates of change for increasing the number of passages under the fixed-testlet-length conditions were similar in real and simulated data analyses.

Condition 3: Manipulating the Scenarios of Items Nested within Passages for a Fixed Total Test Length of 40 Items

Figure 3 show how the test score reliability would change for the 8 possible scenarios of items nested within passages for the fixed total test length of 40 items, when they are estimated through the p x (I:H)) design. The lower solid and dotted lines represent the $E\rho^2$ and the Φ coefficients from the real data analysis, while the upper solid and dotted lines, those from the simulated data.

Insert Figure 3 about here

As shown in Figure 3, the lowest $E\rho^2$ coefficient was obtained for the first nesting scenario in the real data analysis, that is, when all 40 items were generated from a single passage in the test ($40i$:$1p$), whereas the highest $E\rho^2$ coefficient was obtained for the last scenario when there were only one item generated for each of the 40 passage ($1i$:$40p$). When the number of passage was increased from 1 to 4 (while decreasing the number of items per passage from 40 to 10), there was rather a drastic increase in the score reliability. However, the score reliability was impacted only negligibly by adding more passages beyond 4 passages and decreasing the number of items per passage below 10 (*i.e.*, $8i$:$5p$, $5i$:$8p$, $4i$:$10p$, $2i$:$20p$, $1i$:$40p$). In other words, among these 5 scenarios of 5 or more passages for the fixed test length of 40 items, there were not drastic increases or decreases in score reliability. The same trend was observed for the Φ coefficients.

In the simulated data analysis, there was also a slight increase in score reliability, when the number of passages were increased from 1 to 4, but an increase in score reliability was only negligible, compared to that in real data analyses. Beyond 4 passages, the test score reliability remained almost identical. Overall, the trend lines for the simulated (no LID) data were much more flatter than those for the real (LID) data. In other words, the score reliability was not impacted in any meaningful sense, no matter what possible item nesting scenarios were used for the fixed test length in simulated data for the no-LID situation.

Summary and Discussion

The purpose of a study reported in this paper was to examine the impact of passage-related *local item dependence* (LID) and the number of passages and items per passage on test score reliability of an EFL reading comprehension test from the perspective of *generalizability* (G) *theory*. The validity of the testlet-based reliability estimation method was investigated in comparison to the G-theory-based approach as well.

Impact of Passage-Related LID on Score Reliability

The item α coefficients for different test lengths from the p x I design were higher than the passage α coefficients from the p x T design and the $E\rho^2$ coefficients from the p x (I:H)) design in the real data analysis. In simulated data generated with no LID assumed, the item α was also slightly higher than the passage α and the $E\rho^2$ coefficients, but the difference between these coefficients was so small, compared to that in the real data analysis. It may provide strong support for a claim that the item α method would overestimate test score reliability, when it is applied to a test consisting of testlets containing LID items. In other words, the difference between the item α and the $E\rho^2$ coefficient (or passage α) in the real data analysis might be an inflated portion of the item α due to LID. Even though the inflated portion in the item α (about .02 for the 40-item test) seems small in its absolute value, more than 8 items (or 2 passages) need to be added to the current test (*i.e.*, increasing test length more than 20%) to recover the lost portion of alpha due to LID, if the Spearman-Brown prophecy formula (Henning, 1987) is applied.

Another important observation was that the passage α from the p x I design and the $E\rho^2$ coefficients from the p x (I:H) design were very close and almost identical in the real data analysis. It might basically demonstrate the validity of the testlet-based reliability estimation method used for reading comprehension tests made up of passage-based testlets containing LID items. Theoretically speaking, the $E\rho^2$ coefficient from the p x (i:t) design can be regarded as a more ideal way to compute the score reliability of a test made up of passage-based testlets than the testlet-based method, because the item and passage effects are separately modeled and used in the computation of test score reliability (Lee, Brennan & Frisbie, 2001; Lee & Frisbie, 1999). However, in this study, the testlet-based method performed equally well compared to the G-theory-based method.

Impact of Passages and Items on Score Reliability

The impact of increasing the number of passages were similar across the real and simulated data sets, regardless of what "fixed testlet length" conditions were adopted in the analysis. To put it another way, the rates of change in score reliability for increasing the number of passages were uniform across the LID and no LID situations. What was really intriguing was, however, that the impact of adding more items per passage seemed

to have differential impact on the test score reliability in the LID and no LID situations. In Figure 2, the dotted line based on simulated (no LID) data was consistently higher than the solid line based on the real (LID) data for the 3 "multiple items per passage" scenarios. Moreover, the difference between the two lines (*i.e.*, a lost portion in reliability due to LID with passages) became larger, as the number of items nested within passages increased.

Such a trend does make a sense and is consistent with our general intuition about LID items. In a sense, the two LID items are somewhat redundant and provide less information than they appear to (Chen & Thissen, 1997). Since the two LID items are not worth being counted as two items due to a certain amount of redundancy between the items (probably should be counted as something worth between 1 and 2 items), increasing the number of LID items may not have as much impact on the score reliability as increasing the truly independent items. As more LID items are added to the passage, the damaging impact on score reliability could be accumulated over items. This may explain why the lost portion of score reliability increased, as the number of items per passage increased. In other words, there would be less gain in score reliability for increasing the number of LID items than the non-LID items in the passage.

On a related thread, it was also found that score reliability could be maximized by increasing the number of passage than the number of items per passage for the fixed total test length. The highest score reliability was obtained, when there was only one item per passage, whereas the lowest score reliability was obtained, when all 40 items were generated from a single item. As the number of items per passage increased for a fixed test length, the score reliability became smaller. It can be inferred that increasing the number of passages is a more effective way to maximize the score reliability than the increasing the number of items per passage for a fixed total test length, when there is LID present within passages. In a stark contrast, the score reliability was impacted only negligibly (or virtually not!), no matter what item nesting scenarios were used for the same fixed test length in the simulated data analysis. It means that, if all of the test items in the test are truly locally independent, the same level of score reliability should be obtained, whatever item-passage-nested scenarios are employed.

Reliability versus Validity Trade-Off

Passage-based item sets have been the important building blocks of item writing and test construction in reading comprehension assessment, with many strengths and advantages. There may be a trade-off, however, between "reliability" and "validity" in using them in the test, which may be analogous to the conflict between "internal consistency reliability" and "content validity" often mentioned in the literature (Henning, 1987). Content validity in part may largely depend on the diversity and comprehensiveness of items sampled and included in the test, whereas internal consistency reliability may rely on homogeneity of items included in the test. In the case of passage-based reading assessment, a wider range of comprehension sub-skills and processes could possibly be captured in the test more easily, when multiple questions are

asked of the examinees about their understanding of the content of the passage from various perspectives. Such a format may also enhance the "utility" aspect of validity in test use (Messick, 1989), because it could allow test developers and administrators to design and administer the tests within the reasonable limits of development costs, testing time, and other resources.

On the other hand, the sharing of a passage among item is very likely to create potential redundancy and dependency among items within passages and a subsequent reduction in test score reliability due to LID. As the results of the real and simulated analysis show, given the same test lengths of 40 items, a test made up of locally-independent items can achieve a higher score reliability than the one made up of item sets containing LID items. Moreover, it was also found that the most efficient measurement scenario for maximizing the score reliability of the fixed test length would be to use a "one-item-per-passage" strategy, when the test score reliability alone is put in focus. In a sense, such a strategy would enable us to sample and include a wider array of passage topics and content areas, even though it may be an unfeasible option for many practical testing situations. However, it should be noted that the reliability is not the only important psychometric factor that should be considered in determining the final configuration and format of the test. In many practical testing situations, it would be very unimaginable to adopt the "single-item-per-passage" strategy, requiring examinees to read 40 texts to answer 40 questions (only a single item per text), for instance, not to mention of time and resources needed to prepare 40 adequate passages for each form of the test. Moreover, it might also force item writers in the end to use only shorter texts to reduce development cost and testing time, which might impact validity and authenticity of EFL reading comprehension assessment for intermediate and advanced EFL learners.

All these taken together, the real issue might be how to realize the full potential of passage-based item sets in terms of authenticity and utility without seriously sacrificing the score reliability of the test. One possible option might be then to keep the current *passage-based item-set* format for the authenticity and utility reasons and find a way to minimize redundancy and LID among items within a passage despite the sharing of a common passage among items. A series of investigation may be necessary to identify and understand the item and passage content features that might exacerbate or reduce passage-related LID on both theoretical and empirical bases, even though such a goal may be hard to accomplish in a short period of time. Once these factors are identified and understood well, such new knowledge can be incorporated into item analysis, review and revision as well as initial item writing. Accumulated research findings on this line of research can inform item writers and test developers so that they might create updated guidelines for item writing, review, and analysis.

In addition to this line of research efforts, a graph like Figure 3 can be used first to find several equally-good measurement scenarios of items nested within passage that could achieve an pre-set or acceptable level of test score reliability. Then, such information can be combined with other content constraints including the feasibility and

availability of required resources in determining the final configuration of the reading comprehension test made up of passage-based items sets.

Future Research and Conclusion

Avenues for Further Investigation

This study has demonstrated that the G-theory approach can be effectively used to investigate the impact of LID on test score reliability and determine the most optimal measurement scenarios for passage-based item sets. Further investigations on the following two areas would prove very valuable in understanding the nature of the test consisting of *passage-based testlets*:

First, the reading test was specially designed in this study in such a way that the same number of items (*i.e.*, 4 items per passage) is given in each of the 10 passage, which made it possible to use the balanced design in the G-study to estimate the variance components. In practice, the assumption of a balanced design is rarely met for the reading comprehension tests, because the different numbers of items are created for each passage depending upon the lengths and content of the passage. Lee and Frisbie (1999) have demonstrated that the same p x (I:H) adopted in this study can be also used for the unbalanced design, with some modification. Further research on the test score reliability for the unbalanced design would makes it easy to apply the G-theory-based approach in the practical testing situations for EFL reading assessment.

Second, the interplay between LID and text and item characteristics in the passage-based testlets needs to be better understood (Wainer & Wang, 2000). The degree of redundancy in the passage could be a potential factor impacting the overall level of LID for a passage-based testlet. In relation to the item characteristics, Muraki and Lee (2001) has found through content analysis of an extreme LID item pair that the key option of one question made all of the distracters candidates for a correct option for another question within the same passage. Lee (*in press*) has also reported that information overlap in the phrasing of the key option for one question and the item stem in another question in the same passage could exacerbate LID between a pair item sharing a passage. It may indicate that when the same informational element in the passage helps examinees to answer both questions, a level of LID between the items can increase. Further investigation on this line might be able to help us to better understand the relationships between the text characteristics and passage-related LID.

Conclusion

Passage-based item sets have been important building blocks of item writing and test construction in reading comprehension assessment, but the contemporary item–based psychometrics have created a mismatch between test construction and analysis due to their inability to model passage-related LID within passages. G-theory now makes it possible to separately model the passage and item effects in the estimation of score

reliability for the reading comprehension tests made up of passage-based item sets, solving this mismatch. The findings of this study shed important lights on the impact of passage-related LID on score reliability, the validity of the testlet-based reliability estimation (passage α) method, and the optimal strategies for maximizing the score reliability in a reading comprehension test composed of passage-based item sets. These findings can surely pave a promising road ahead for realizing the full potential of passage-based item sets in enhancing the authenticity and validity of EFL reading comprehension assessment.

References

Alderson, C. (2000). *Assessing reading*. Cambridge University Press.

Andrich, D. 1985: A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. Embretson (Ed.), *Test design: Contributions from psychology, education, and psychometrics (pp. 245-273)*. New York: Academic Press.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: ACT.

Chen, W., and Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Educational and Behavioral Statistics, 22, 265-289.*

Crick, G. E. & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system (ACT Tech. Bulletin No. 43)*. Iowa City, IA: American College Testing Program.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cureton, E. E. 1965: Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25*, 327-346.

Ferrera, S., Huynh, H. & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Assessment, 36*, 119-140.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 12*, 21-25.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Henning, G. (1987). *A guide to language testing: Development, evaluation, and research*. Boston, MA: Hienle & Hienle.

Henning, G. (1989). Meaning and implications of the principle of local independence. *Language Testing, 6*, 95-108.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2,* 76 - 104.

Lee, G., Brennan, R. L. & Frisbie, D. A. (2001). Incorporating the testlet concept in testlet score analyses. *Educational Measurement: Issues and Practices, 19*(4), 9-15.

Lee, G. & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12(3)*, 237-255.

Lee, Y. (in press). Examining passage-related local item dependence (LID) and measurement construct using Q_3 statistics in an EFL reading comprehension test. *Language Testing*.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Mehrens, A. W. & Lehmann, I. J. (1978). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart and Wilson.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3^{rd} *ed.*) (pp.13-103). New York: American Council on Education/Macmillan.

Muraki, E. (2000). *RESGEN manual: Item response generator* (Version 4.0). Princeton, NJ: ETS.

Muraki, E. & Lee. Y. (2001). *Detecting local item dependency in the TOEFL reading comprehension section: An application of the full-information item factor analysis*. ETS internal document. Princeton, NJ: ETS.

Sireci, S., Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247.

Suen, H. K. 1990: *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (1991). *MULTILOG user's guide*. Chicago: Scientific Software.

Thissen, D., Steinberg, L. & Mooney J. A. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H. (1996). How is reliability related the quality of test scores? What is the effect of local dependence on reliability. *Educational Measurement: Issues and Practice*, 15 *(1)*, 22-29.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H. & Lewis, C. (1990). Toward a psychometric for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer. H. & Lukhele, R. (1997). *How reliable is the TOEFL test?* TOEFL Technical Report 12 (TR-12). Princeton, NJ: ETS.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22-29.

Wainer, H. & Wang, X. 2000: Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220.

Wilson, M. & Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Author Note

This paper derives from the author's Ph. D. dissertation. Needless to say, the responsibility for any errors that remain are solely mine, and the ideas and opinions expressed in the paper are those of the author, not necessarily of Educational Testing Service (ETS).

Table 1. Estimated Generalizability Coefficients ($E\rho^2$) and Dependability Indices (Φ) for Different Test Lengths from 3 Different D-Study Designs

(a) Estimated $E\rho^2$ and Φ Coefficients for Real Data (n_p =1072)

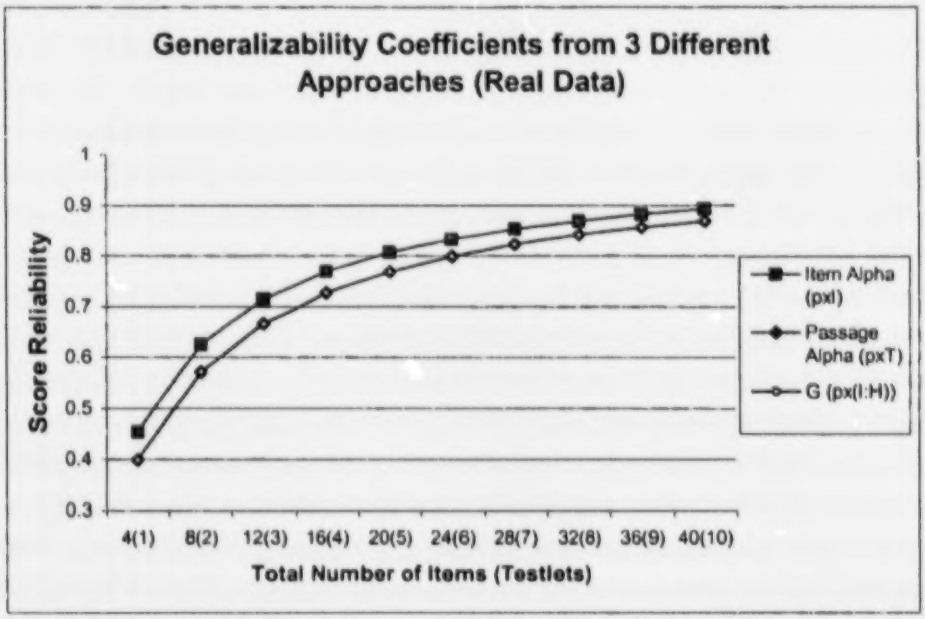
No. of Passages	No. of Items	D-Study Designs					
		p x I		*p* x T		*p* x (I:H)	
		$E\rho^2$ (Item α)	Φ	$E\rho^2$ (Passage α)	Φ	$E\rho^2$	Φ
1	4	0.45	0.44	0.40	0.40	0.40	0.39
2	8	0.63	0.61	0.57	0.57	0.57	0.56
3	12	0.71	0.70	0.67	0.66	0.67	0.65
4	16	0.77	0.76	0.73	0.72	0.73	0.71
5	20	0.81	0.80	0.77	0.77	0.77	0.76
6	24	0.83	0.82	0.80	0.80	0.80	0.79
7	28	0.85	0.85	0.82	0.82	0.82	0.81
8	32	0.87	0.86	0.84	0.84	0.84	0.83
9	36	0.88	0.88	0.86	0.86	0.86	0.85
10	40	0.89	0.89	0.87	0.87	0.87	0.86

(b) Estimated $E\rho^2$ and Φ Coefficients for Simulated Data (n_p =1072)

No. of Passages	No. of Items	D-Study Designs					
		p x I		*p* x T		*p* x (I:H)	
		$E\rho^2$ (Item α)	Φ	$E\rho^2$ (Passage α)	Φ	$E\rho^2$	Φ
1	4	0.45	0.43	0.44	0.43	0.44	0.42
2	8	0.62	0.60	0.61	0.60	0.61	0.59
3	12	0.71	0.69	0.70	0.70	0.70	0.68
4	16	0.76	0.75	0.76	0.75	0.76	0.74
5	20	0.80	0.79	0.79	0.79	0.79	0.78
6	24	0.83	0.82	0.82	0.82	0.82	0.81
7	28	0.85	0.84	0.84	0.84	0.84	0.83
8	32	0.87	0.86	0.86	0.86	0.86	0.85
9	36	0.88	0.87	0.87	0.87	0.87	0.86
10	40	0.89	0.88	0.89	0.88	0.89	0.88

Figure 1. Generalizability Coefficients ($E\rho^2$) from 3 Different Reliability Estimation Approaches (p x I, p x T, p x (I:H)).

(a) Real Data (n_p=1072)



(b) Simulated Data (n_p=1072)

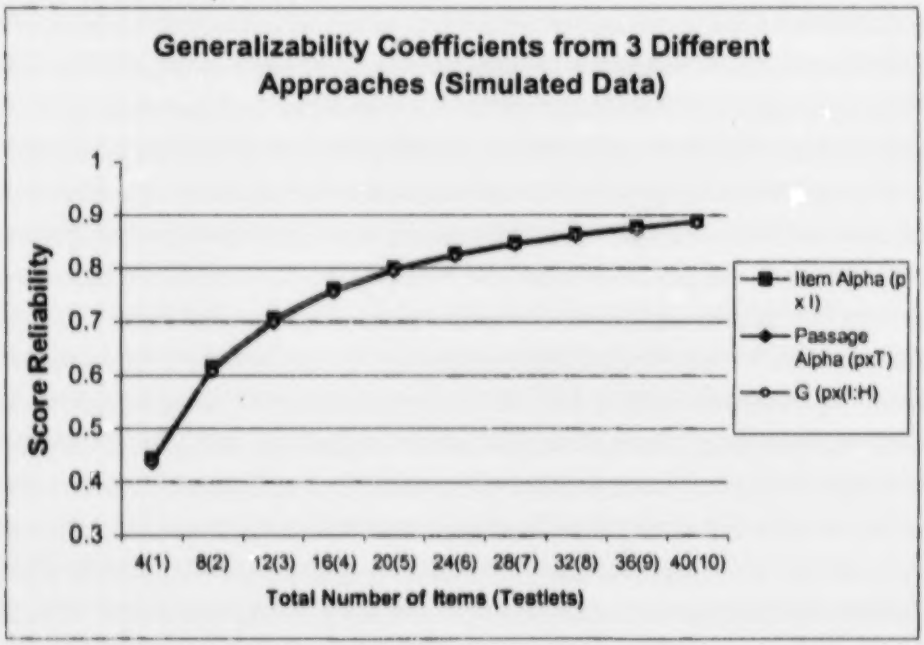


Figure 2. Generalizability Coefficients for Different Number of Passages with the Number of Items Nested Within Passages Fixed to 1, 2, 3, and 4 (p x (I:H))

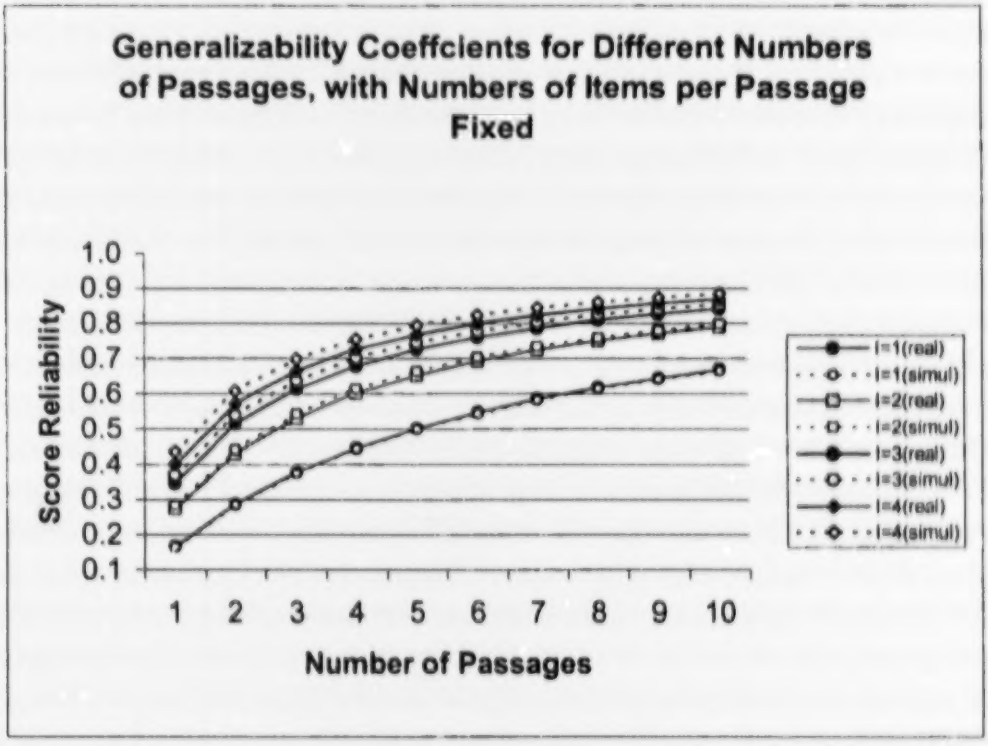
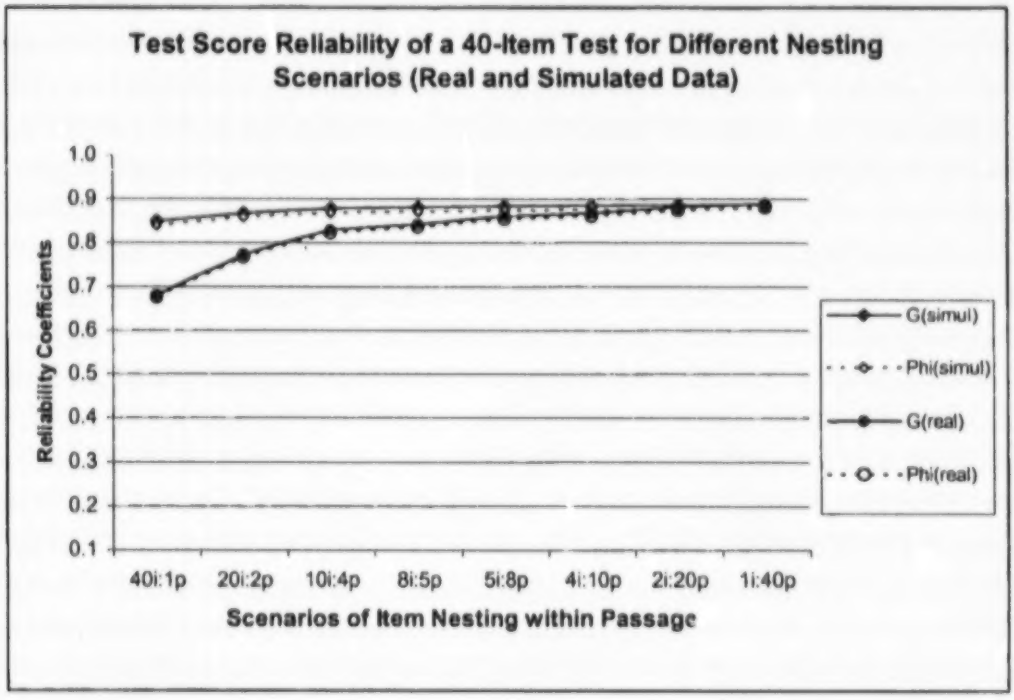


Figure 3. Generalizability and Dependability Coefficients for Various Scenarios of Items Nested Within Passages (p x (I:H)) for the Fixed Test Length of 40 Items



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM034449

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Score Reliability of a Test Composed of Passage-Based Testlets: A Generalizability Theory Perspective	
Author(s): Yong-Won Lee	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☐	Level 2A ☒	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Yong-Won Lee	Printed Name/Position/Title: Associate Research Scientist
Organization/Address: Educational Testing Service MS 19-R ETS Rosedale Road Princeton, NJ 08541	Telephone: 1-609-734-1348 FAX: 1-609-734-1755
	E-Mail Address: ylee@ets.org Date: 9/12/2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)